Supporting Information

Insights into global diatom distribution and diversity in the world's ocean

Shruti Malviya, Eleonora Scalco, Stéphane Audic, Flora Vincent, Alaguraj Veluchamy, Julie Poulain, Patrick Wincker, Daniele Iudicone, Colomban de Vargas, Lucie Bittner, Adriana Zingone, Chris Bowler

This file includes:

- Materials and Methods
- Figures S1 to S8
- Dataset 1 and 2

SI Materials and Methods

Distance based Analysis

The PR2 database v99 (1) contains 2,947 full length 18S unique diatom sequences. These sequences were aligned and sequence variations along the entire sequence were used to define the hypervariable regions. Entropy calculation was done on all reference sequences (Fig. S1A). Pairwise distances were calculated for the full length and all hypervariable regions using Kimura-2-parameter model (2). V4 and V9 sequences were used to check the performance in differentiating the four prominent phylogenetic clades of diatoms, i.e., radial centric, polar centric, araphid pennate and raphid pennate. Each of the V4 and V9 hypervariable regions and full-length 18S rDNA sequences were aligned using MUSCLE and phylogenetic inference was done with NJ algorithm using pairwise distances in MEGA6 (2). The tree was statistically tested using 1000 bootstraps. A reference database was obtained and all the reference sequences were aligned. Shorter sequences (less than 125 nucleotides) along with extremities were eliminated to obtain the same sequence lengths. To evaluate the ability of the V9 region to differentiate between the intragenus and intergenus variation among diatom V9 sequences, we calculated p-distance between all pairs of reference sequences (Fig. S1).

Extracting diatom V9 metabarcodes from global eukaryotic protistan metabarcoding data set

A total of ~580 million quality-checked reads, representing ~2.3 million unique rDNA ribotypes (V9 region of 18S rDNA), were generated from 334 photic-zone plankton communities sampled during the *Tara* Oceans expedition (3). The *Tara* Oceans expedition (4-5) covered seven oceanographic provinces, i.e., North Atlantic Ocean (NAO), Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), and South Pacific Ocean (SPO). At each station, plankton communities were obtained for four size fractions from two water-column depths (SRF and DCM). Total nucleic acids (DNA + RNA) were extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified (3). The V9 reads were quality checked and to reduce the influence of PCR and sequencing errors, only sequences seen in at least two different samples with at least 3 copies were retained. The sequences have been deposited in GenBank (see [4], Accession: PRJEB4352; ID: 213098). Taxonomy assignments for all ribotypes were obtained through annotation against an expert-curated V9 reference database (for details, see [3]) using the global alignment search strategy implemented in the ggsearch36 program (*Fasta* package). This reference database contains sequences from both cultured strains and the environment, and contained 1,232 unique diatom V9 reference sequences corresponding to 159 genera, with most genera being represented by more than one sequence (Fig. S2). Of the 159 genera in the reference data set, we retrieved 87 genera in our data set. However, only 79 out of 87 were assigned at an identity greater than 85% and were selected for further analysis.

These unique barcodes were taxonomically assigned to known eukaryotic entities based on the PR2 database (1). From this, metabarcodes assigned to diatoms, at a percentage identity of ≥ 85% to the reference sequence, were selected. When BLAST results gave rise to more than one unique best hit, a last common taxonomy of the best BLAST hits was created [3]. Moreover, in order to improve the assignation of the barcodes that couldn't be assigned to the genus level, as the PR2 database version we used was based on a former release of Genbank [3], PR2 assignations were also compared to Genbank assignations (release 210 from October 2015). We manually checked each assignation, and kept the best of two (PR2 or Genbank assignation) based on the best percentage identity value and BLAST scores. We could thus improve our conclusions and assignations, in particular when working with sequences that could not be assigned to the genus level (SI Dataset 2).

All the barcodes were clustered into biologically meaningful operational taxonomic units (OTUs) using the 'Swarm' approach (6). This method uses 1 base pair difference (local threshold) between barcodes. It also overcomes input-order dependency induced by centroid selection, a typical bias of classical clustering methods (6).

Morphological analyses

The 20-180 µm size fraction samples selected for microscopy analyses included SRF and DCM samples from the Cape Agulhas region (Stations 52, 64, 65, 66, 67, and 68), the South Atlantic transect (Stations 70, 72, 76, and 78), the Southern Ocean stations (Stations 82, 84 and 85), and South Pacific Ocean stations (Stations 122, 123, 124, and 125). Three ml of each sample was placed in an Utermöhl chamber with a drop of calcofluor dye (1:100,000), which stains cellulose thus allowing to better detect and identify diatom species. Cells falling in 2 or 4 transects of the chamber were identified and enumerated. Phytoplankton species were identified and enumerated using a light inverted microscopy (Carl Zeiss Axiophot200) at 400x magnification. The identification was performed at the species level when possible.

Reassignment of unknown diatom ribotypes

Four diatom V9 rDNA ribotypes were chosen (marked with asterisk (*) in Fig. 8, *lower panel*) for reassignment, based on their presence in the top 20 most abundant unassigned diatom ribotypes in the whole *Tara* metabarcode dataset. The goal was to amplify a longer 18s rDNA fragment of the target diatom from 18s rDNA preamplified samples in order to improve the quality of the sequence taxonomy. Preamplification of 18s rDNA was performed on DNA extraction of ethanol fixed sea water collected for each of the *Tara* metabarcoding samples (TV9_172, TV9_225, TV9_361 and TV9_339). DNA was extracted with MasterPureTM DNA/RNA purification kit (Epicenter) and PCR amplified using the universal-eukaryotic primers (forward Euk-A [5'-aacctggttgatcctgccagt-3'] and reverse Euk-B [5'-tgatcctcctgcaggttcacctac-3']) from Medlin et al. (7). Amplifications were performed with the Phusion™ high-fidelity DNA polymerase (Finnzymes) in a 50-µL reaction volume, using the following PCR parameters: 30 s at 98 °C; followed by 15 cycles of 10 s denaturation at 98 °C, 30 s annealing at 57.5 °C, and 30 s extension at 72 °C; with a final elongation step of 10 min at 72 °C. PCR product was purified with Nucleospin® PCR Clean-up (Macherey-Nagel).

For each target diatom ribotype, the equivalent 18s rDNA preamplified sample in which its relative abundance was the highest was chosen for PCR (Polymerase Chain Reaction), in order to maximise chances of amplifying the ribotype with highly specific reverse primers.

The forward primer chosen was the D512F (D512F: 5'-ccgcgtaattccagctccaatagcg-3') universal diatom primer from Zimmerman et al. (8). The reverse primer was designed in order to find the 3’ end consecutive eight base pairs 100% specific to the target sequence that matched the lowest number of non-specific sequences in the sample. Four ribotypes and their respective reverse primer sequences are listed below:

Sample ID	Ribotype md5sum ID	Reverse primer sequences
TV9_172	01eb4d181204cc0e142f55f1632b0b8c	172_rev: 5’-aggttcggacaagttctcgcggtcag-3’
TV9_225	4d2d2df1f3cdb2080ace0b23c17928be	225_rev: 5’-ttcctactaaatgataaggtttagacgagt -3’
TV9_361	ba6c7a54f4f24e0888797d4e062cda61	361_rev: 5’-ggggacaagttctcgcggctaacat-3’
TV9_339	8e6521a0e8234f3660e5c0d302c33da9	339_rev: 5’-gcggagacaagttctcgcgacagat-3’

TV9_179: St82|0.8-inf|srf; TV9_225: St85|20-180|srf; TV9_361: St123|5-20|srf; TV9_339: S122|5-20|dcm

The unassigned V9 sequence was cut in windows of 8 base pairs and each of them was mapped against all positions of the OTU sequences present in the sample under Perl 5 (version 16, subversion 2 (v5.16.2)). A heatmap of hits was obtained, giving the number of times each window perfectly matched a position in the V9 sequences of the sample. The best primer candidates were then extended to 26 base pairs on average, and the final primer was chosen based on its position in the target sequence, close to the end of the V9, its GC content, Tm and checked on IDTDNA Oligo Analyzer 3.1 (http://eu.idtdna.com/calc/analyzer). Temperature gradient PCR from 58 to 68 °C were performed to obtain highest specificity of the primers to the target DNA.

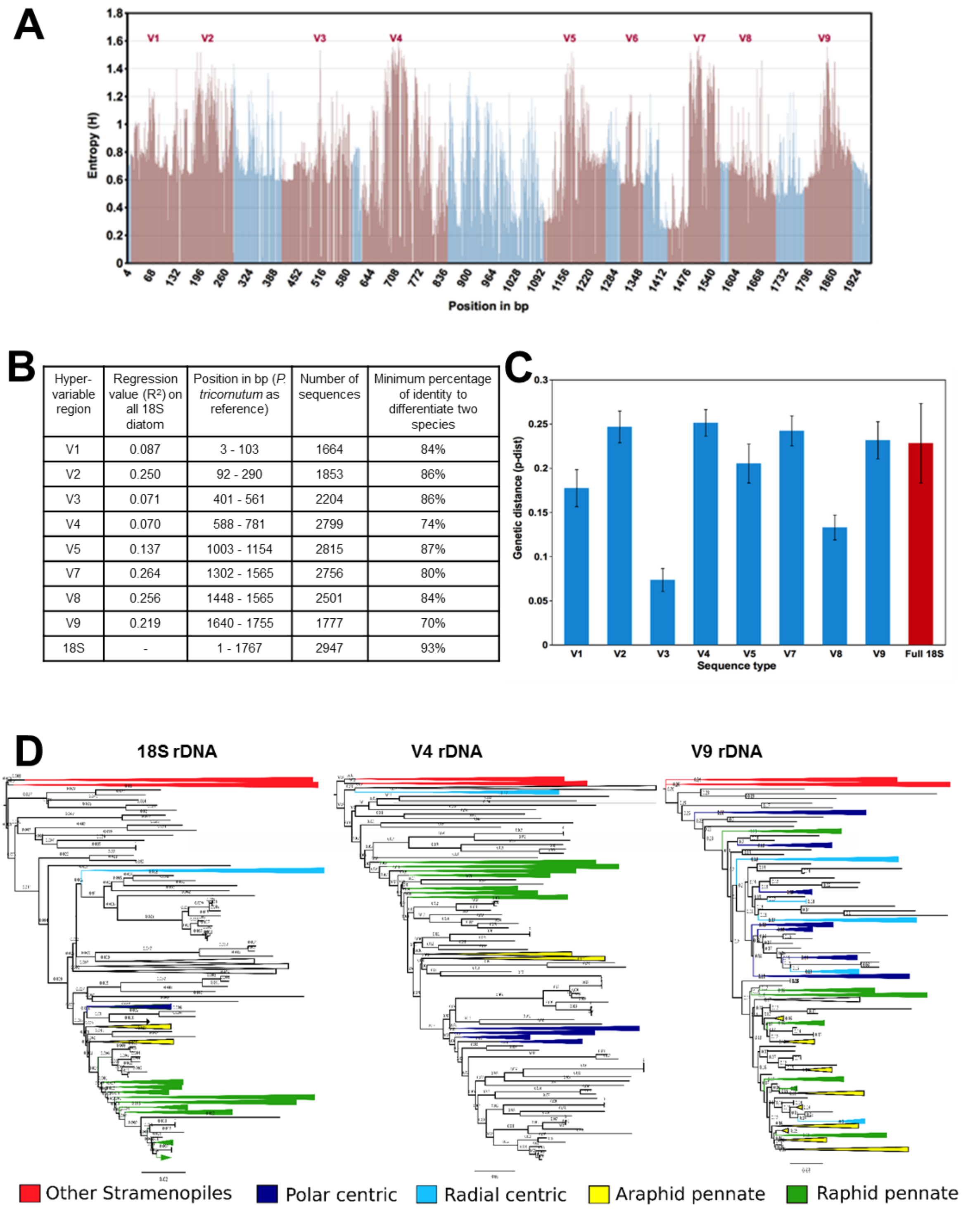
Amplifications were performed with the Phusion™ High-Fidelity DNA Polymerase (Thermo-Scientific™) in a 20-µL reaction volume, using the following PCR parameters: 30 s at 98 °C; followed by 33 cycles of 10 s denaturation at 98 °C, 30 s annealing from 66 to 68 degrees, and 60 s extension at 72 °C; with a final elongation step of 10 min at 72 °C. DNA was extracted from agarose gel with Nucleospin® Gel and PCR Clean-up (Macherey-Nagel) and directly sent to GATC Biotech for paired-end Sanger Sequencing. Resulting sequences were assigned by blastn in NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

References

1. Guillou L, et al. (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 41(D1):D597-D604
2. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* 30:2725-2729.
3. de Vargas C, et al. (2015) Eukaryotic plankton diversity in the sunlit global ocean. *Science* 348(6237):1261605.
4. Pesant S, et al. (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* 2:150023.
5. Karsenti E, et al. (2011) A holistic approach to marine eco-systems biology. *PLoS Biol* 9:e1001177.
6. Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2:e593.
7. Medlin L, Elwood HJ, Stickel S, Sogin ML (1988) The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* 71:491-499.
8. Zimmerman J, Jahn R, Gemeinholzer B (2011) Barcoding diatoms: evaluation of the V4

subregion on the 18S rRNA gene, including new primers and protocols. *Org Divers Evol* 11(3):173-192.

9. Leblanc K, et al. (2012) A global diatom database - abundance, biovolume and biomass in the world ocean. *Earth Syst Sci Data* 4:149-165.
10. OBIS (2015) Data from the Ocean Biogeographic Information System. Intergovernmental Oceanographic Commission of UNESCO. Web. http://www.iobis.org (consulted on 2015/07/29).



Hyper-variable region	Regression value (R^2) on all 18S diatom	Position in bp (*P. tricornutum* as reference)	Number of sequences	Minimum percentage of identity to differentiate two species
V1	0.087	3 - 103	1664	84%
V2	0.250	92 - 290	1853	86%
V3	0.071	401 - 561	2204	86%
V4	0.070	588 - 781	2799	74%
V5	0.137	1003 - 1154	2815	87%
V7	0.264	1302 - 1565	2756	80%
V8	0.256	1448 - 1565	2501	84%
V9	0.219	1640 - 1755	1777	70%
18S	-	1 - 1767	2947	93%

Fig. S1. Assessing V9 hypervariable sub-sequence (130 bp) of small-subunit (SSU) ribosomal RNA (rRNA) genes as diversity markers. (A) 2,947 full-length 18S rDNA sequences were obtained from the PR2 reference database corresponding to 718 diatom species (1). They were aligned and entropy along the full length was computed. The sequence variations along the entire length was used to assess the nine hypervariable regions (V1-V9) using the *RNAstructure* program. Regions in red are V1-V9. The bases are numbered according to the alignment position. (B-C) Hypervariable region performance against the 18S rDNA sequence. Length variation and pairwise genetic distances calculated using the Kimura-2-parameter model for all nine hypervariable regions are shown. Regression of V1-V9 p-distance by NJ on to that of the 18S

sequence shows that V5 could best explain the phylogeny, followed by V4. Although the mean genetic distances were better in V4 and V9, they may not explain the phylogeny well. The performance of V9 was 23 % less than that of the full-length 18S sequence, and taxa assignment at less than 70 % identity in the V9 region was found to be insufficient for diatoms. (D) Phylogenetic inference on Bacillariophyta from full-length 18S rDNA sequence phylogeny, V4 rDNA phylogeny, and V9 rDNA based phylogeny rooted at Stramenopiles. Four prominent phylogenetic clades of diatoms, i.e., radial centric, polar centric, araphid pennate, raphid pennate are known. V4 and V9 sequences were used to check their performance in differentiating these four groups. Each of the hypervariable regions and full-length 18S rDNA sequences were aligned using MUSCLE and phylogenetic inference was done with NJ algorithm using pairwise distances in MEGA6. The tree was statistically tested using 1,000 bootstraps. We found that the V4 region wrongly placed some raphid pennate diatoms within centric groups, whereas the V9 region could not differentiate well between radial and polar centric diatoms, nor between raphid and araphid pennate groups (Fig. S1D).

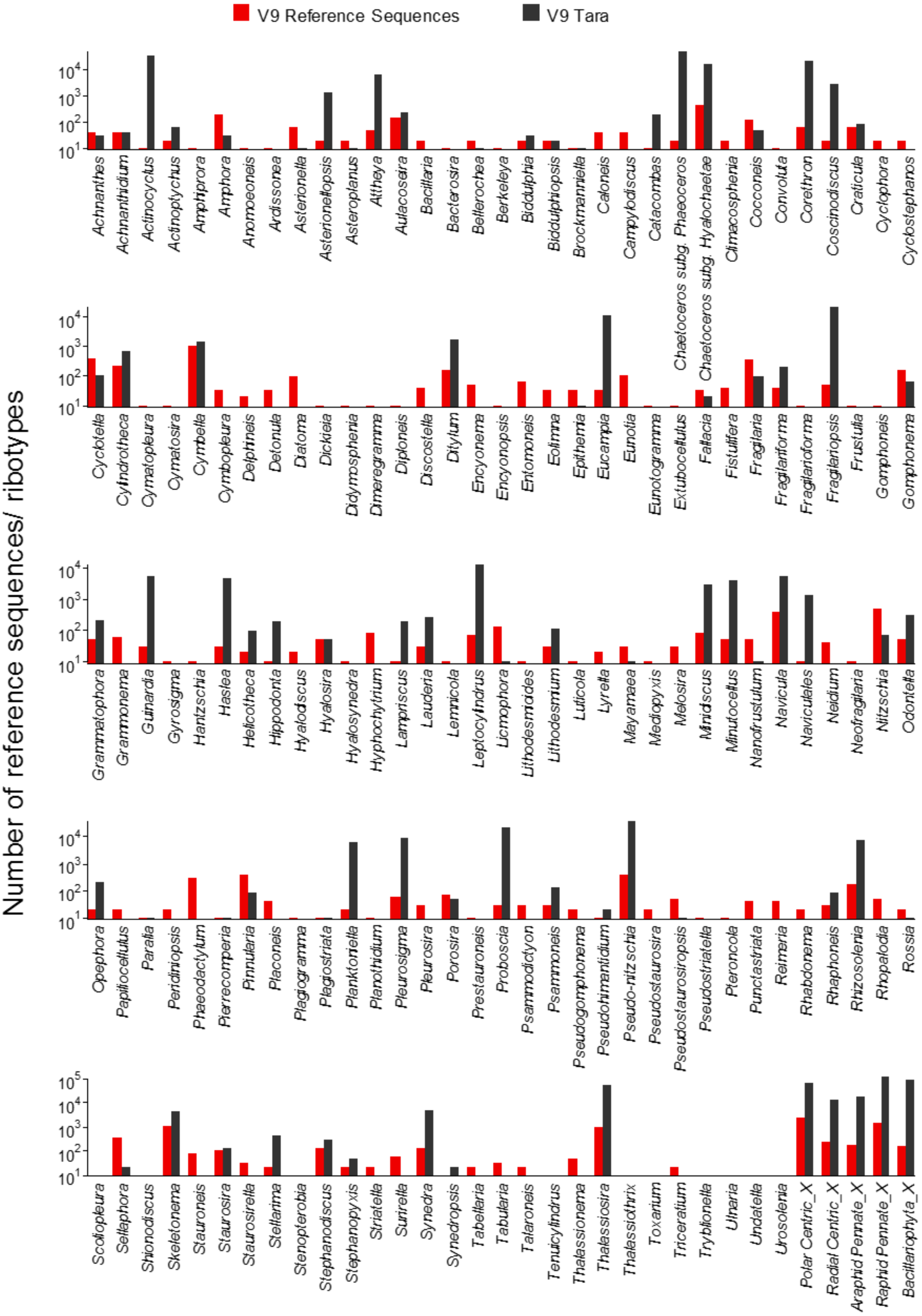


Fig. S2. Novelty in *Tara* Oceans diatom ribotype data set. Barplot showing the number of reference sequences present for each genus and the total number of unique V9 tags from *Tara* Oceans data set assigned to it. The reference database has a total of 1,648 V9 sequences annotated as being derived from diatoms. The level of percentage identity to the reference sequence varied across ribotypes, but for this analysis a similarity cut-off of 85% was used. From a total of 63,371 ribotypes, 30,057 ribotypes were unassigned due to the lack of reference sequence.

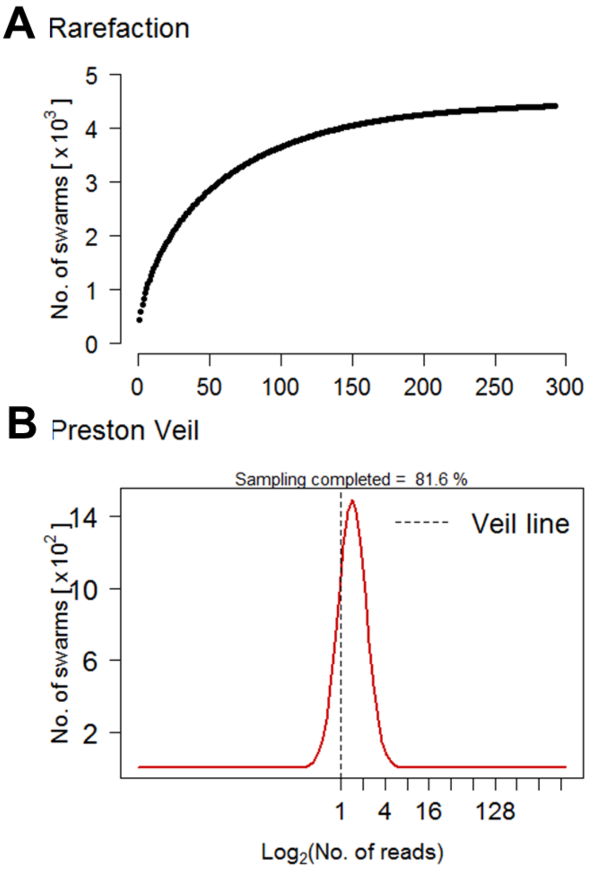
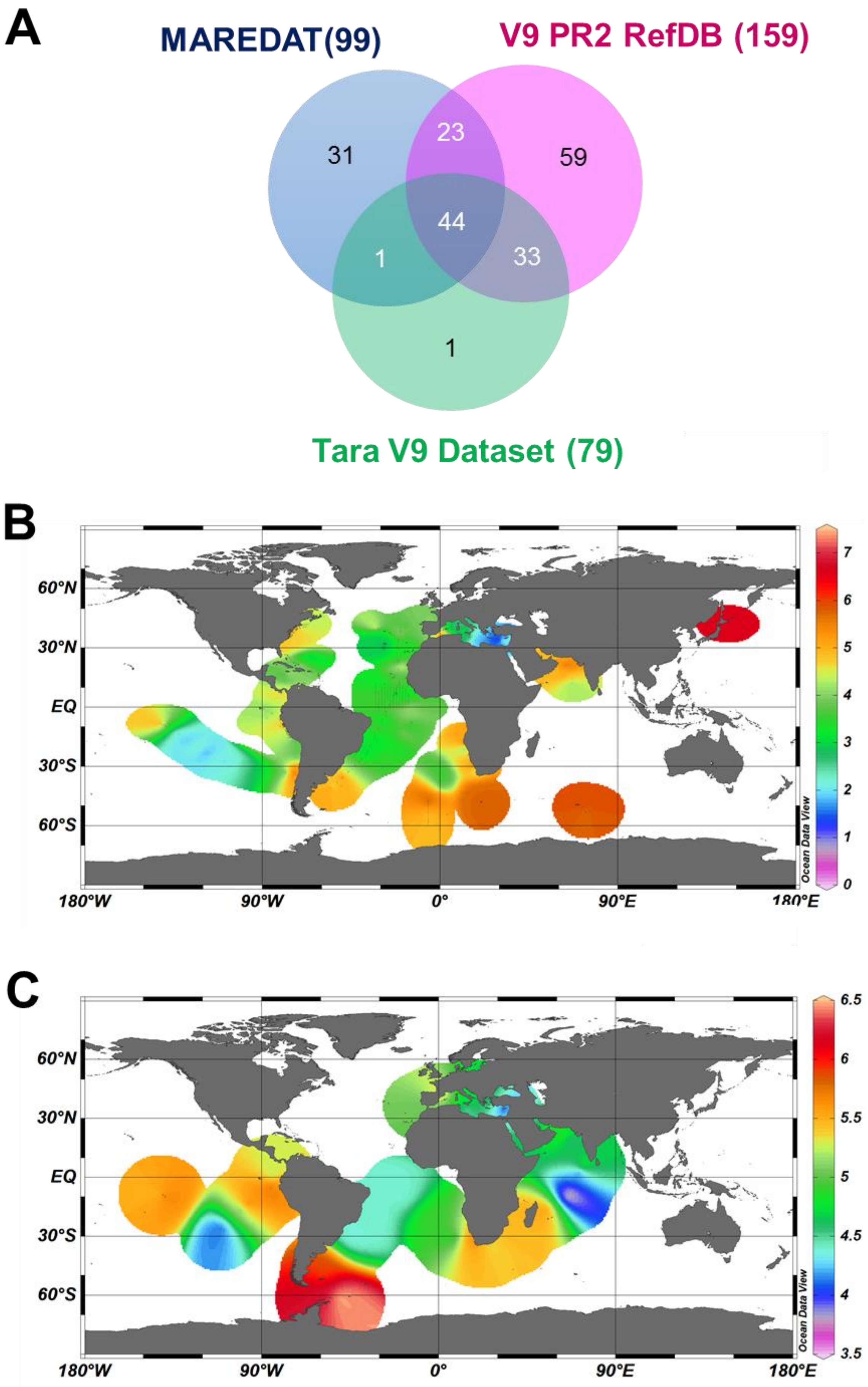


Fig. S3. Completeness of the diatom ribotype data set based on OTUs. (A) OTU rarefaction curve. A sample-based rarefaction curve, representing OTU richness for diatoms. (B) Estimating the completeness of sampling based on OTUs. OTU abundances were log$_2$-transformed. Most of them were seen with intermediate abundances with a relatively few rare or very few ubiquitous OTUs. The area under the Preston curve provides an extrapolated estimate of richness and thus an indication of the completeness in the sampling effort. The theoretical OTU richness inferred from Preston Veil was found to be 4,748, indicating 873 OTUs undetected. OTU calling was based on Mahé et al. (6).


A
MAREDAT(99)
V9 PR2 RefDB (159)
31
23
59
44
1
33
1
Tara V9 Dataset (79)
B
60°N
30°N
EQ
30°S
60°S
180°W
90°W
0°
90°E
180°E
7
6
5
4
3
2
1
0
Ocean Data View
C
60°N
30°N
EQ
30°S
60°S
180°W
90°W
0°
90°E
180°E
6.5
6
5.5
5
4.5
4
3.5
Ocean Data View

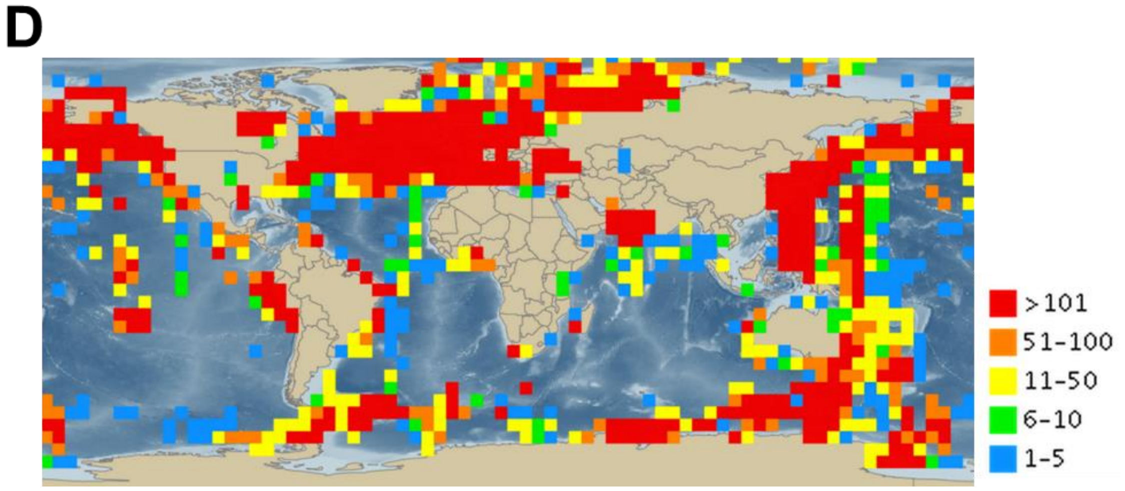


Fig. S4. Comparing diatom distributions and abundance obtained from our study to the reports from MAREDAT and OBIS. (A) Venn diagram showing the overlap between *Tara* Oceans data set, the V9 PR2 reference database, and MAREDAT. The green circle represents the subset of reference genera identified in the *Tara* Oceans data set. (B) Distribution of Bacillariophyta obtained from MAREDAT database (9). The colour represents log-transformed cell counts per litre. (C) Distribution of Bacillariophyta obtained from *Tara* Oceans data set. The colour represents log-transformed total V9 ribotype abundance. (D) Global abundance of Bacillariophyta species obtained from OBIS datasets (10) (each square is coloured according to the abundance of diatoms species observed in the area).

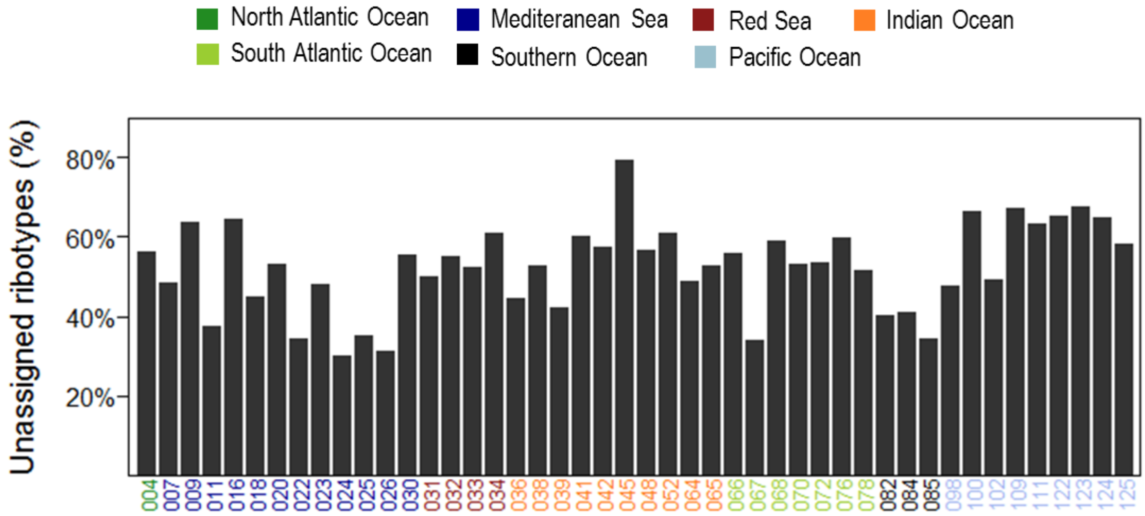
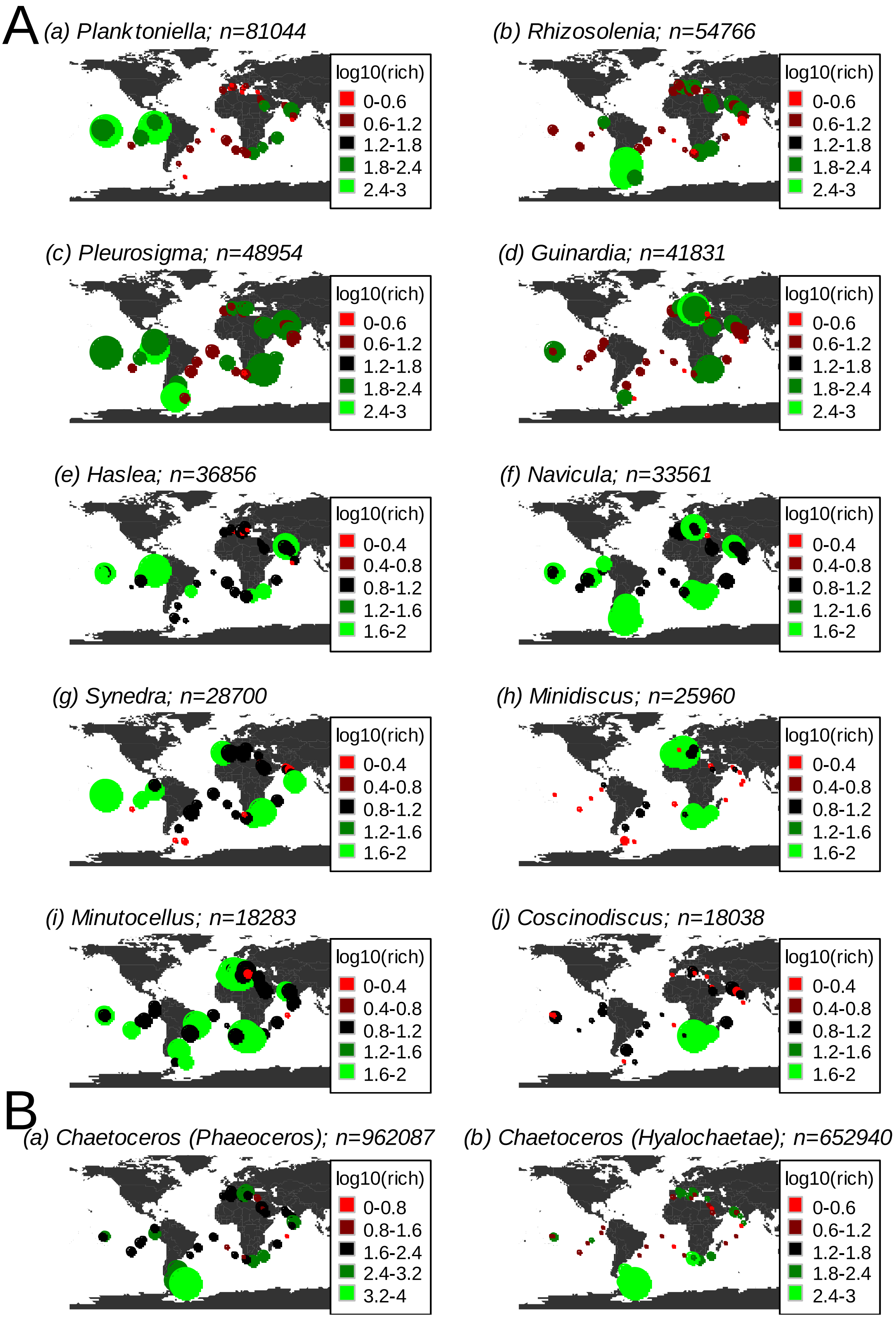


Fig. S5. Percentage of unassigned ribotypes in each station. Within each station, 31-81 % of the ribotypes could not be assigned to any known diatom genera. The highest proportion of unassigned ribotypes was seen in Station 45 (~79 %) followed by stations in the Pacific Ocean (Stations 109,111,122,123,124) (~58-68 %). Stations with the highest abundance (Stations 67 and 85) contained ~34 % of unassigned ribotypes.



A (a) *Planktoniella; n=81044*

(b) *Rhizosolenia; n=54766*

(c) *Pleurosigma; n=48954*

(d) *Guinardia; n=41831*

(e) *Haslea; n=36856*

(f) *Navicula; n=33561*

(g) *Synedra; n=28700*

(h) *Minidiscus; n=25960*

(i) *Minutocellus; n=18283*

(j) *Coscinodiscus; n=18038*

B (a) *Chaetoceros (Phaeoceros); n=962087*

(b) *Chaetoceros (Hyalochaetae); n=652940*

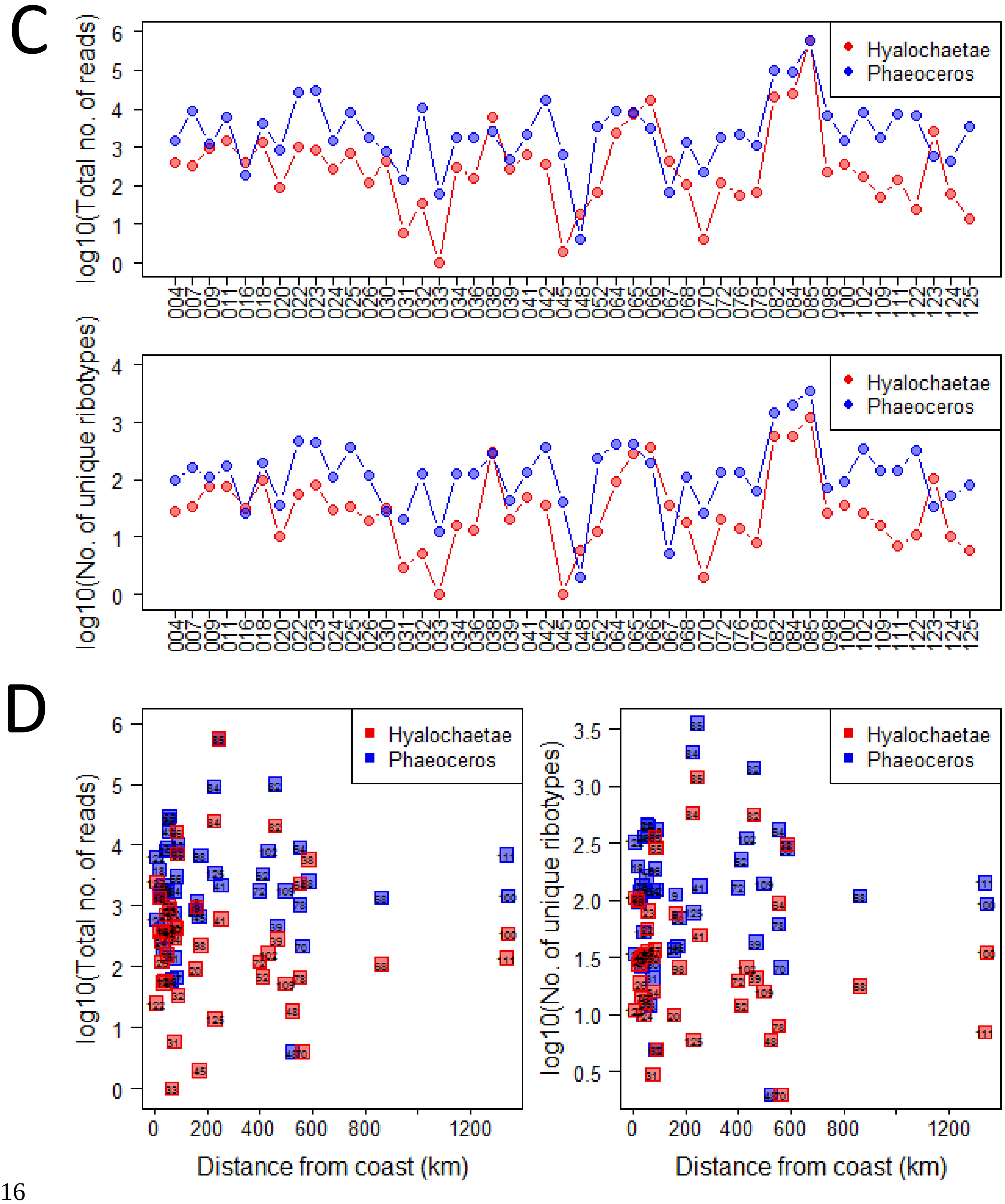


Fig. S6. Global distribution and diversity of abundant genera. (A) Genera ranked 11 to 20 based on their ribotype abundance. (B) *Chaetoceros (Phaeoceros)* and *Chaetoceros (Hyalochaetae)*. (A & B). The area of the circle is scaled to the total number of reads for each genus at each location. For each panel, the color key is shown in the legend. Red - low richness; Green - high richness. (C) Abundance of the two *Chaetoceros* subgenera in different stations. (D) Abundance in different stations of each *Chaetoceros* subgenus compared to distance from the coast. Each square corresponds to a station. Left panel was built using reads and right panel was built using ribotypes. Red dots/squares correspond to *Hyalochaetae* and blue dots/squares correspond to *Phaeoceros* data.

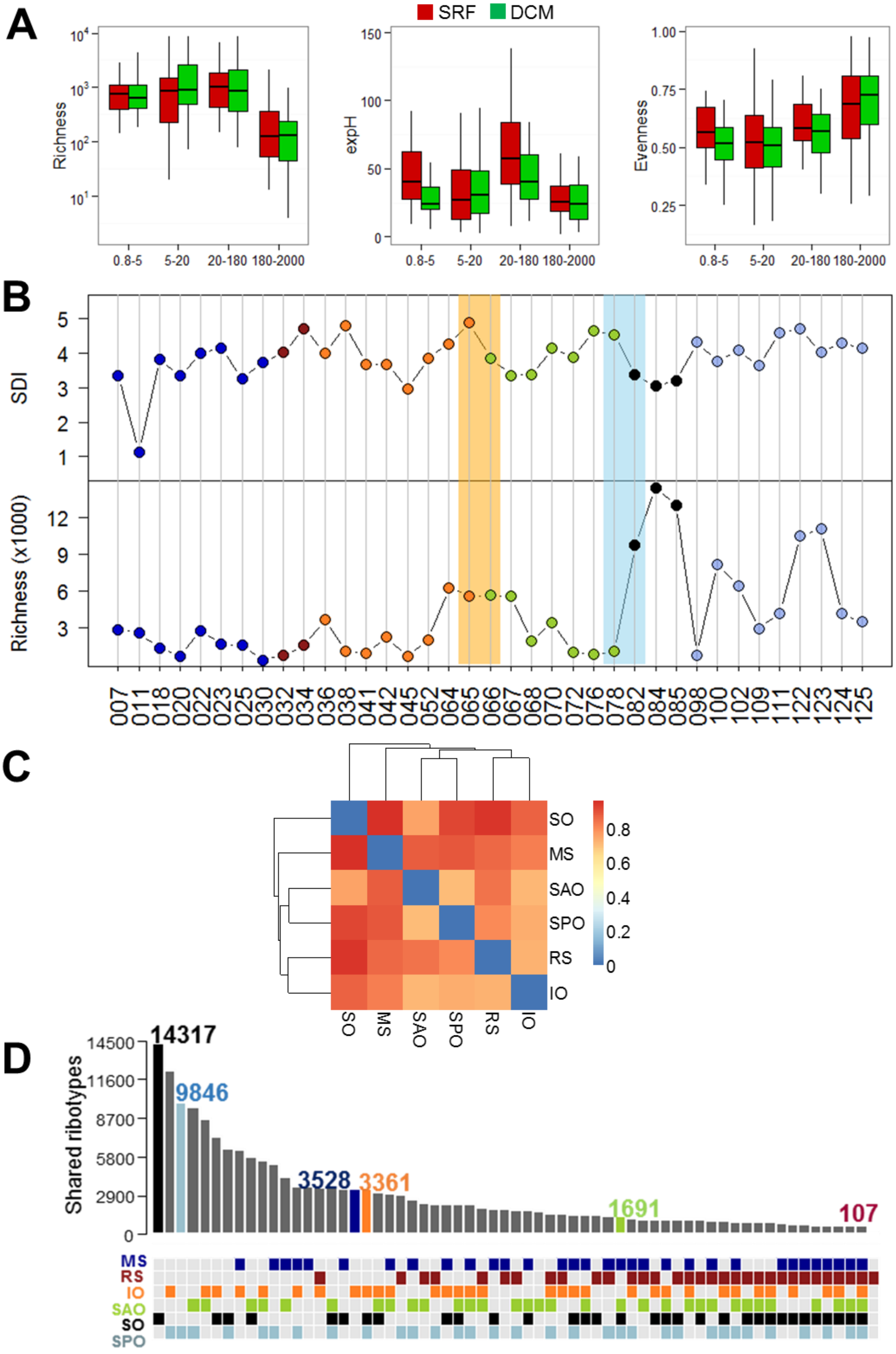

A
SRF
DCM
Richness
10^1
10^2
10^3
10^4
expH
0
50
100
150
Evenness
0.25
0.50
0.75
1.00
0.8-5
5-20
20-180
180-2000
B
SDI
1
2
3
4
5
Richness (x1000)
3
6
9
12
007 011 018 020 022 023 025 030 032 034 036 038 041 042 045 052 064 065 066 067 068 070 072 076 078 082 084 085 098 100 102 109 111 122 123 124 125
C
SO
MS
SAO
SPO
RS
IO
0
0.2
0.4
0.6
0.8
D
Shared ribotypes
0
2900
5800
8700
11600
14500
14317
9846
3528
3361
1691
107
MS
RS
IO
SAO
SO
SPO

Fig. S7. Variations in diatom diversity. (A) Ribotype richness, effective number of species (expressed as expH) and evenness are shown per depth and size-class. The results indicate that the 20-180 µm size fraction was the most diverse and showed highest diversity at DCM. The smaller 0.8-5 µm size fraction also showed similar trends. The largest size fraction exhibited the lowest abundance and richness but has the highest evenness among all size fractions. In general, SRF samples were found to be less diverse and even than DCM samples. (B) Variation in diatom diversity across stations. Spatial variation of diatom diversity across 37 stations inferred from Shannon Diversity Index (SDI) and richness. (C) Pairwise community dissimilarity (Bray-Curtis) across provinces, signifying higher dissimilarities for higher values. (D) Shared number of ribotypes among oceans. Bar graph showing the number of ribotypes shared between oceanic provinces; presented from left to right, from greatest to least number of shared ribotypes. Counts are based on presence-absence. The color-coded numbers above the bars indicate the ribotypes exclusive to each province.

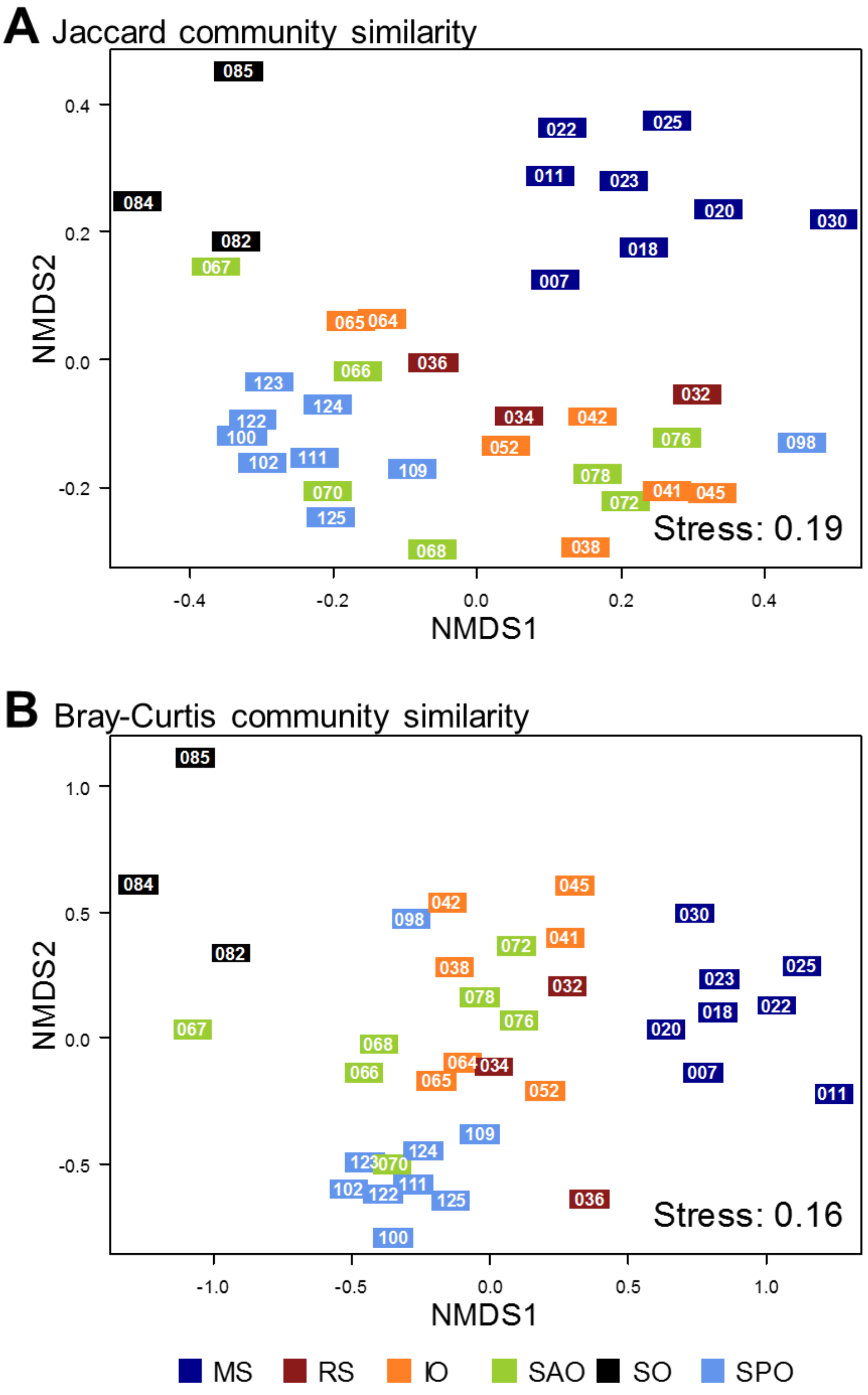


Fig. S8. Diatom community composition. (A) Incidence measure. Pairwise Jaccard dissimilarity was used to cluster stations hierarchically (group-average linkage). A two-dimensional NMDS ordination indicated that communities grouped according to oceanic provinces, albeit with significant overlapping (stress=0.19). Each of these oceanic clusters were significantly different (ANOSIM; R = 0.58; W = 0.001). (B) Abundance-based measure. The two-dimensional NMDS ordination of the transformed data in reduced space with a stress value of 0.16 was used to visualize pairwise Bray-Curtis distance among stations. Hellinger transformation was performed on the abundance matrix to minimize the influence of rare ribotypes. Each symbol corresponds to a station, colored based on oceanic province.